Introduction

Every forty seconds someone in the world intentionally take their own life. This claim stems from the World Health Organization (WHO) national suicide prevention report. This amounts to about eight hundred thousand suicide per-year and the second leading cause of death among 15–29-year-olds globally. The Veteran Affairs National Suicide Data Report shows suicide rates have increased by 29.5 percent for US Veterans and 20.6 percent for non-Veteran adults between 2005 and 2016. This epidemic is a tragedy that occurs in every corner of the world, leaving countless loved ones, family members, friends, and comrades heartbroken. The objective of this case study is to identify and evaluate possible relationships between suicide rate and country-based measures between 2005 and 2016. Ultimately, exploring the underlying effects of intentional self-harm mortality at the international level is the first step towards global suicide prevention.

Data Summary

The final data set analyzed in this study was comprised of three sources representing two governing organizations, the World Health Organization and the World Bank Group. Google's BigQuery Application programming interface (API) was implemented through kaggle.com's Python type kernel to find, collect, and merge twenty country-based measures; for which eleven variables from the World Development Indicators Data (WDI) and two variables from the World Bank: Education Data (World Bank) contributed to the case study. WHO's Statistics data visualizations dashboard was used to generate and download a data table containing the response variable, suicide rate, and another potential regressor. Refer to Table 1 in the appendix for descriptions and source of the response variable and the fourteen potential predictors. In total, there were 180 observations, each representing a unique country in 2010. Refer to Table 2 in the appendix for summary statistics of each variable.

Methods and Results

The data collection and preparation process involved twenty-three queries and four file merges. The creation of a Python dictionary and user defended R function was necessary to uniquely identify each country and ensure proper merging of data. In addition to varying country names, some sources of data identified regions or groups of counties within their records. These errors were not apparent and resulted in the repeated analysis of the data as discrepancies where identified and corrected.

The initial analyzation of data was conducted through basic descriptive summaries of each variable. Five variables that originated from the World Bank: Education Data source where removed due to an overwhelming amount of missing values. These variables were not included in the introduction nor Table 1. The model selection process started by comparing an exhaustive search method with the forward stepwise selection. Results from the exhaustive search method showed the adjusted r-squared value increasing from about 0.18 for a single regressor of percent population growth to about 0.34 for a fit with seven predictors. While the CP value did not seem to converge on the expectation, p, but seeming avoided that value. The AIC from the first iteration in the forward stepwise method showed an almost equivalent value for eight of the fourteen regressors, which was above the null model's AIC. The addition of percent population growth in the following step caused all but two regressors to fall below the null's AIC. After the stepwise method added those regressors, internet users, and the probability of dying, unemployment female decreased in AIC by 20 and proceeded to be selected over the null model. Consequently, unemployment and unemployment male decreased in AIC below the null model. The forward stepwise method was concluded after the addition of unemployment. This result of suicide rate given growth, internet users, probability of dying, unemployment female, and unemployment, seemed to be a reasonable best fit in the exhaustive search result. Where the model without unemployment female and unemployment had slightly less desirable measures of adjusted r-squared, CP, and BIC.

Looking at scatter plots and correlations between suicide rate, growth, internet users, probability of dying, unemployment female, and unemployment reviled a strong positive correlation between both unemployment measures. Incrementally removing each unemployment measure revealed that their coefficients significance, at the 0.05 level, was dependent on both being considered in the model. Therefore, the model selected for residual analysis was the suicide rate given growth, internet users, and the probability of dying. The estimated coefficients and related results of this fit are shown in Table 3.

The diagnostic plot of Figure 1 shows a possible violation in the assumed linear relationship between suicide rate and the regressors, shown in the residuals vs fitted subplot. More notably, this subplot shows a gross violation in non-constant variance of the residual values with a cone or funneling shape observed as the fitted values increase. The Normal QQ subplot of Figure 1 indicates a heavy-tailed distribution of the residuals compared to the assumed normal distribution. Also, there is a distinct and possibly extreme observation identified in the Residual vs Leverage plot values of this model. However, it is well below a generally accepted cooks-D value of one. Multicollinearity was evaluated for this model fit by calculating the VIF values for each regressor. Internet users per 100 people had a VIF of 1.7, the highest of the three regressors which are well below a generally accepted limit of 5.

Attempts to address violations of non-constant variance and non-normal residual distribution were done by transforming the response, suicide rate. Table 4 compares diagnostic measurers of the non-transformed fit of suicide rate to the log and square root transformations. The adjusted r-squared had decreased by about 10 and 3 percent for the log and square root transformations. However, no estimated coefficients became non-significant for ether transformation. Figure 2 compares the diagnostic plots, column-wise, for the standard fit, log-transformed, and square root fits respectfully. Looking at subplot (1,2) and (1,3) of the Residuals vs Fitted values for the log and square root transformations, we see that the log transformation reduced a large majority of non-constant variance

with exception of a region of negative residuals for fitted values just above 2. This would explain why the adjusted r-squared associated with this log transformation had a larger decrease compared to the square root. Summarily, this long downward tail has a large impact on the normality of the residuals. As far as the square root transformation, there seem to be small improvements in non-constant variance, even seeming to correct the normality in the residuals but this predominate tail still exists. We cannot reasonably conclude that the log or square root transformation has corrected any violations.

Weighted least squares were applied to both of the transformed models in an attempt to address the non-constant variance in this off-centered region of fitted values. Slight improvements where observed when the square root method was used on the fitted values to finds weights but still beyond reasonable to calm that the assumptions of non-constant variance where meet by ether weighted, transformed fit.

Conclusions

Exploring the underlying effects of intentional self-harm mortality across 180 countries is far from complete. Within this case study, underlining multicollinearity has limited the Extortionary Least Squares Method in explaining the variation in suicide rate given many potential regressors collected. However, hope is far from lost for the 200 thousand humans that succumb to suicide each year. Now that methods have been developed to combined multiple sources of country-based measures, we can continue the search for potential regressors. One question, in particular, is worth asking in light of this analysis; what underlining relation exists between percent population growth and suicide rate? As I alluded to in the introduction, this country-level investigation of suicide is mortality a first step in the understanding and ultimate reduction of self-harm mortality. We must not lose perspective of this crippling and merciless epidemic by comparing across countries. Any successful effort at saving our brothers and sister must include a personal understanding. This deep personal aspect of suicide is

difficult to infer at this level of comparison. Yet, I still believe there is much to gain from refining a worldwide approach to suicide prevention.

Considerations for revision must address the short-comings we observed within this study. The World Health Organization has shown that 2016 had disproportionate suicide rates for low income or under developing countries. Implementing indicator variables may improve our association with this noted trend. Also, in considering our sample size of 180 countries analyzed here; implementing a random sampling method may revive trends that were washed out or hidden in the large volume of data.

APPENDIX: Tables and Figures

Table 1

Variable	Unit	Source	Definition
suicide_rate	People per 100,000 population	WHO	Weighted average of suicide deaths from the WHO Global Health Estimates (GHE) 2016, using population estimates produced by the UN Population Division.
dying	Percent	WHO	Percent of 30-year-old-people who would die before their 70th birthday from any of cardiovascular disease, cancer, diabetes, or chronic respiratory disease, assuming that s/he would experience current mortality rates at every age and s/he would not die from any other cause of death
growth	Percent	World Bank	Exponential rate of growth of midyear population from year t-1 to t, expressed as a percentage.
GDP	U.S. dollars	WDI	GDP at purchaser's prices is the sum of gross value added by all resident producers in the economy plus any product taxes and minus any subsidies not included in the value of the products.
pop	Count	WDI	Total population is based on the de facto definition of population, which counts all residents regardless of legal status or citizenship. The values shown are midyear estimates.
female	Count	WDI	Female population is based on the de facto definition of population, which counts all female residents regardless of legal status or citizenship.
male	Count	WDI	Male population is based on the de facto definition of population, which counts all male residents regardless of legal status or citizenship.
density	People per sq. km of land area	World Bank	Population density (people per sq. km of land area)
Unemployment	Percent	WDI	Unemployment refers to the share of the labor force that is without work but available for and seeking employment.
UnemployMale	Percent	WDI	Unemployment, male (% of male labor force) (modeled ILO estimate)
UnemployFemale	Percent	WDI	Unemployment, female (% of male labor force) (modeled ILO estimate)
tourism	U.S. dollars	WDI	International tourism, receipts for travel items
cell	Count per 100 population	WDI	Mobile cellular subscriptions (per 100 people)
netUsers	Count per 100 population	WDI	Internet users (per 100 people)
air	Count	WDI	Air transport, passengers carried

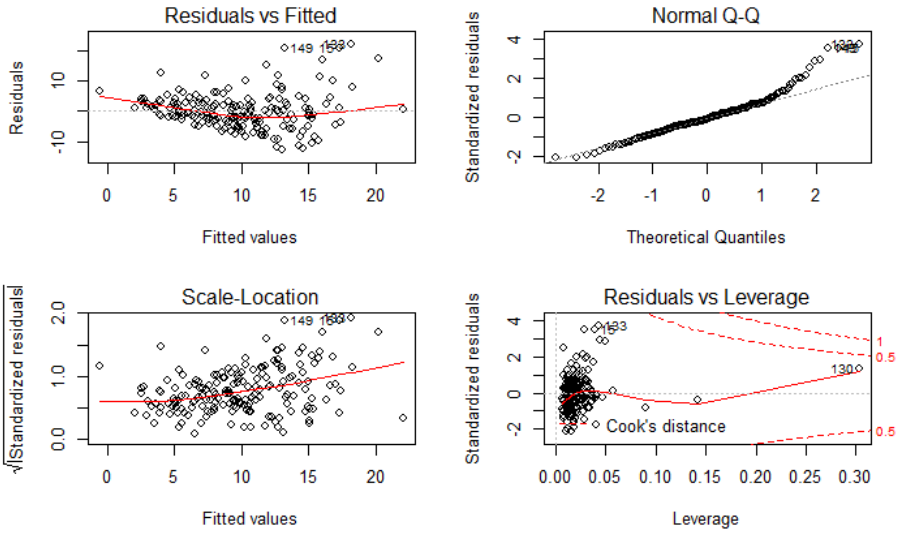
Table 2

Variable	n	Mean	SD
suicide_rate	180	9.89	7.20
dying	180	20.12	5.75
growth	180	1.54	1.56
GDP	180	358601981111.49	1338106151816.71
pop	180	38009646.20	139432593.08
female	180	18855880.06	67629990.58
male	180	19177206.98	71811092.55
density	180	177.95	572.06
Unemployment	180	8.17	6.11
UnemployMale	180	7.52	5.81
UnemployFemale	180	9.61	7.57
tourism	180	4872771666.67	13151158166.74
cell	180	87.48	41.48
netUsers	180	31.88	27.05
air	180	14327390.95	59637497.29

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-1.387	2.71	-0.512	0.609	-6.735	3.961
growth	-1.372	0.309	-4.434	0	-1.982	-0.761
netUsers	0.123	0.022	5.551	0	0.079	0.167
dying	0.47	0.099	4.723	0	0.274	0.666

Results of suicide rate fitted on population growth, internet users, and probability of dying.

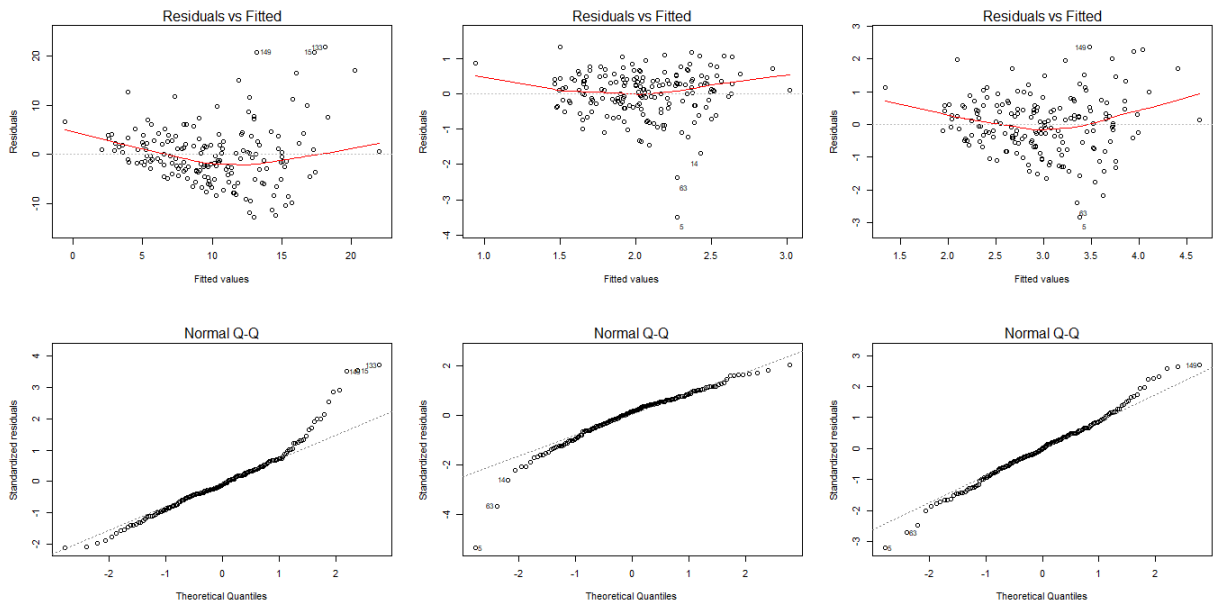
Figure 1



Model	adj.r.squared	sigma	p.value	df	deviance	AIC / BIC
fit_1	0.304	6.007	0	4/176	6351.774	1162.3 / 1178.2
log(y)	0.194	0.653	0	4/176	74.97	363.2 / 379.1
sqrt(y)	0.276	0.888	0	4/176	138.768	474 / 490

Diagnostic table comparing transformed models of suicide rate to the non-transformed fit_1.

Figure 2



Comparing the diagnostic plots by columns for non-transformed, log-transformed, and square root fits respectfully.

Citations

World Health Organization (WHO, 2018). National suicide prevention strategies: progress, examples and indicators. Retrieved from: https://www.who.int/mental_health/suicide-prevention/en/

The World Bank: Education Statistics (2017). Retrieved from: https://www.kaggle.com/theworldbank/world-bank-intl-education

WHO Statistics data visualizations dashboard (2018). Retrieved from: http://apps.who.int/gho/data/node.sdg.3-4-data?lang=en

World Development Indicators Data (WDI, 2016). Retrieved from: https://www.kaggle.com/bigquery/worldbank-wdi

R-Code Attached Separately